Improving Asset Price Prediction When All Models are False

Garland Durham[1] and John Geweke[2]

Revision 2, November 2012

[1]Division of Finance, Leeds School of Business, University of Colorado, USA; Garland.Durham@colorado.edu

[2]Centre for the Study of Choice, University of Technology Sydney, Ultimo NSW 2007, Australia; Erasmus University, The Netherlands; and Division of Finance, Leeds School of Business, University of Colorado, USA; John.Geweke@uts.edu.au. Geweke acknowledges financial support for this work through Australia Research Council grant DP110104732.

Abstract

This study considers three alternative sources of information about volatility potentially useful in predicting daily asset returns: daily returns, intraday returns, and option prices. For each source of information the study begins with several alternative models, and then works from the premise that all of these models are false to construct a single improved predictive distribution for daily S&P 500 index returns. The prediction probabilities of the optimal pool exceed those of the conventional models by as much as 5.29%. The optimal pools place substantial weight on models using each of the three sources of information about volatility.

KEYWORDS: EGARCH, intradaily returns, model combination, optimal pool, S&P 500, stochastic volatility, VIX.

JEL classification: G17 primary, C52 secondary

1 Introduction

Prediction of financial asset prices is important in a variety of private and public sector policy contexts. Examples include the pricing of options by private traders; the measurement of risk in mortgage pools by banks and Federal agencies; and assessment of systemic risk by regulatory agencies and macroeconomic policy makers. In all of these decision-making activities formal prediction models for asset prices are essential tools, and the academic literature has responded with a wide variety of candidates. Yet, those with responsibilities for such decisions recognize that all of these models are incomplete descriptions of reality. How should a decision-maker proceed, knowing that all the models at her disposal are false?

The academic literature provides little practical guidance on this point. The orthodox rational expectations framework is not designed for this purpose. It avoids the issue by assuming that reality is fully described by a parametric model that is known to economic agents and policy makers. When this approach is extended to the situation where economic agents and policy makers must learn about reality, it is typically in the context of a correctly specified parametric model with unknown parameters. The mainstream econometric literature is also unhelpful for the decision-maker confronted with an array of alternative model-based predictive distributions for asset prices. Non-Bayesian econometrics emphasizes specification testing. But when all the available models are false, passing a battery of tests is an indicator of insufficient sample size and test power rather than evidence that a particular model is true and others are false. With sample sizes sufficiently large and tests sufficiently powerful, all models will be rejected, leaving the prediction question unresolved. Bayesian econometrics provides an elegant operational theory of model combination, but because it is founded on the explicit condition that one of the models under consideration is a literal description of reality it shares the limitations of the rational expectations literature.

This paper looks at several different classes of models that generate predictive distributions for asset prices by making use of alternative sources of information (daily returns only, high-frequency intraday returns, and option-implied volatility), and uses the method of

optimal prediction pooling developed in Geweke and Amisano (2011) to construct predictive densities that outperform any of the individual models. The situation is typical in that each class of models provides a distinct window into the underlying reality, but we do not believe any of them to be literally true. The optimal pooling idea makes explicit allowance for the possibility that all of the models under consideration are false and reflects the observed behavior of decision-makers, who are likely to consult several different models when making policy, even models that have been rejected by formal statistical tests. While such behavior seems paradoxical, it is supported by the finding that pools are typically able to outperform even the best of the individual models they encompass, sometimes by a large margin, while placing significant weight on models that are easily rejected by conventional tests.

The study proceeds in two steps. First, we construct the collection of individual models. We look at a total of 42 models categorized into three groups based on the source of information used to forecast volatility. The models allow for flexibility in the shape of the predictive distribution, the way this shape changes over time, and the relationship between observed and latent volatility, building on methods introduced in Durham (2007). The second step utilizes this extended collection of asset price prediction models to generate improved predictive distributions using the method of optimal prediction pooling. The application uses daily S&P 500 index returns from the first trading day of 1990 through the end of March 2010.

The models are described in Section 2. The first group uses the history of daily returns only and comprises six models: two stochastic volatility (SV) models with leverage and four exponential generalized autoregressive heteroscedasticity (EGARCH) models. In the second group, consisting of 18 models, the indicator of daily volatility is the sum of squared 5-minute S&P 500 Index futures returns from previous trading days. The third group, also consisting of 18 models, uses the Chicago Board Options Exchange Market Volatility Index (VIX), which is a model-free indicator of daily volatility constructed from options prices.

In each group of models, we begin with a simple base model and elaborate on it in several

directions. In all groups the daily return shock is either Gaussian or a mixture of two normal distributions. In all models except the stochastic volatility models the volatility component is either a single factor (the conventional treatment) or the sum of two independent factors with different autocorrelation properties. Permitting more than one volatility factor introduces flexibility into the autocorrelation of the latent volatility state, allowing the models to generate, for example, long memory-like behavior. In the high-frequency and options groups there is a third dimension of flexibility related to the form of the mapping from the volatility state to the scaling factor that determines the conditional variance of daily returns.

Turning to the second step, Section 3 briefly summarizes the essentials of optimal linear prediction pools introduced in Geweke and Amisano (2011). It introduces an intuitive foundation for prediction pools, taking as the point of departure optimal asset portfolios. The optimal linear prediction pool is that linear combination of model predictive densities that, historically, achieves the best outcome using a log scoring rule. Pooling does not invoke the assumption that one of the models under consideration corresponds to the data generating process, but if that is in fact the case then in large samples the optimal linear pool is the data generating process. Otherwise optimal linear pools typically put positive weight on several of the candidate models.

Formal Bayesian procedures, generally known as Bayesian model averaging, also lead to linear prediction pools. They do so under the explicit assumption that one of the models under consideration corresponds to the data generating process. If this assumption is correct then the Bayesian model average coincides with the data generating process in large samples—the same outcome as with an optimal linear pool. If this assumption is incorrect, Bayesian model averaging still leads to a pool consisting of a single model asymptotically, the one for which the directed Kullback-Leibler distance from the data generating process is smallest.

A finding that in large samples the optimal linear pool includes several models with positive weights is evidence against the specification that one of the models is the data generating

process. In these circumstances the optimal linear pool provides predictions superior to those of Bayesian model averaging, as assessed by a log scoring rule. Optimal model pooling does not explicitly specify a set of models that must include the data generating process and is fundamentally a non-Bayesian procedure. As pointed out in Geweke and Amisano (2011, Theorem 6) there are Bayesian mixture models that by construction perform at least as well as optimal linear pools in large samples. But the comparison is asymptotic, and the analytical and computational demands of these models are high. In contrast optimal linear pooling is a simple procedure that often improves substantially on the predictive performance of both individual models and Bayesian model averaging.

Empirical results are reported in Section 4. For the simplest models in each group there are very substantial differences in log scores across the three classes (Section 4.2), with the basic daily model outperforming the basic options model which in turn outperforms the basic high-frequency model. The various model extensions described in Section 2 eliminate the bulk of the differences in log scores among the best models in each group, with the best high-frequency model followed by the best daily model followed by the best options model.

Conventional model combination procedures, motivated by Bayesian model averaging and described in Section 4.3, amount to "winner takes all": on most trading days predictions are based almost entirely on one group of models, but there are sharp fluctuations between groups. In the early part of the sample, options models dominate with occasional reversals in favor of daily models. In the latter part of the sample, high-frequency models mostly dominate. The performance of this model averaging procedure is poor, both in comparison with the best of our extended models and with some simple benchmark predictive density combinations.

Optimal prediction pools, constructed in Section 4.4, behave very differently. Following initial fluctuations, weights in the optimal pools stabilize several years into the sample. By the end of the sample, the high-frequency and daily model groups take on weights of about 0.40 each, with the remainder falling on options models. A related measure of model value

indicates that in the latter years of the sample high-frequency models have the most value followed closely by the daily models. The value of the options models is small but positive. The optimal pools substantially outperform all of the individual models in log score, and they also outperform the simple benchmark predictive density combinations.

This study concentrates on the specific problem of extending and combining models that use alternative sources of information about volatility for the purpose of improving the one-step-ahead prediction of an index of asset prices. For sake of transparency we do not introduce notation or techniques more general than required to address this particular task. Yet the methodology in the study can be extended to a much wider set of similar problems. Some of these extensions are quite modest while others require addressing additional technical issues. Section 5 summarizes the findings of this study and then briefly discusses a much larger set of prediction problems amenable to similar treatment and the work involved.

2 Models and estimation techniques

We look at several classes of models, corresponding to different sources of information about volatility: daily returns, high-frequency, and options. For all of the models, returns are of the form

$$y_t = \mu_Y + \sigma_t \varepsilon_t, \tag{1}$$

where y_t is the daily log return, σ_t is the volatility scaling factor and ε_t is a mixture of normals standardized to have mean zero and variance 1. The model classes differ in the information used to estimate σ_t. In each class, we examine a hierarchy of models with varying amounts of flexibility in several relevant dimensions. All models are estimated by maximum likelihood. Predictive densities are then formed by replacing unknown parameters with their point estimates. In all cases, predictive densities for the time t return are constructed using only information available at time $t-1$.

Our objective is not just to obtain forecasts that match the observed data in, say, first or second moments. Rather, the object of interest is the full predictive density, with assessment using a likelihood-based metric closely related to Kullback-Leibler distance. Thus it is important for the models to have sufficient flexibility to generate realistic distributions, motivating the use of the mixture models. Given enough components these models are able to fit any distribution arbitrarily closely (McLachlan and Peel, 2000). For distributions encountered in applications similar to the one in this paper, good fits are typically obtained with a small number of components.

These mixture models are closely related to the jump models commonly used in this literature. But, we do not take a stand on the nature of the intradaily price movements: what part is diffusive, what part is due to jumps, and what the characteristics of those jumps are. We are only interested in the shape of the daily return distributions. The mixture distributions are useful for this purpose. See Durham (2007) for additional detail.

We examined mixtures of up to three components. The three-component models perform well in the later part of the sample but have difficulty in the early part, where the quantity of available data is more limited. In full Bayesian estimation, the problems in the early part of the sample could be alleviated by using an appropriate prior. With the maximum likelihood approach used in this paper, an analagous effect could be achieved by adding curvature to the likelihood surface in an ad hoc manner. However, for the application in this paper we restrict attention to models with a maximum of two mixture components.

Some of the models include multiple volatility factors, providing flexibility in the autocorrelation characteristics of the latent volatility state. In models with two factors, for example, one captures a persistent long-term trend in the level of volatility, while the other captures short-term fluctuations around it. Such models are capable of generating long memory-like behavior (Bollerslev and Mikkelsen, 1996).

The class of daily models consists of two stochastic volatility (SV) and four exponential generalized autoregressive heteroscedasticity (EGARCH) models. The SV models are of the

form

$$\begin{aligned} y_t &= \mu_Y + \sigma_Y \exp\left(v_{t-1}/2\right)\varepsilon_t \\ v_t &= \phi v_{t-1} + \sigma_V \eta_t, \end{aligned} \tag{2}$$

where y_t is the log return and v_t is the unobserved volatility state. The volatility innovations are of form $\eta_t = \rho\varepsilon_t + (1-\rho^2)^{1/2} u_t$, where $u_t \sim N(0,1)$ is uncorrelated with ε_t. Thus $\mathrm{E}(\eta_t) = 0$, $\mathrm{var}(\eta_t) = 1$ and $\mathrm{corr}(\eta_t, \varepsilon_t) = \rho$, but because ε_t is non-Gaussian so is η_t. Negative values for ρ capture a leverage effect, whereby negative returns are associated with increased volatility on subsequent days. The nature of the relationship between ε_t and η_t implies that extreme price changes will tend to generate large changes in volatility as well. Estimation is done using the simulated maximum likelihood algorithm and EIS sampler of Richard and Liesenfeld (2006). Predictive densities are formed by integrating across uncertainty in the volatility state. We look at two partcular cases of the SV model: `sv_1` uses a Gaussian distribution for ε_t, and `sv_2` uses a mixture of two normal distributions.

The EGARCH models are of form

$$\begin{aligned} y_t &= \mu_Y + \sigma_Y \exp\left(\sum_{i=1}^{k} v_{it}/2\right)\varepsilon_t \\ v_{i,t+1} &= \alpha_i v_{it} + \beta_i\left(|\varepsilon_t| - (2/\pi)^{1/2}\right) + \gamma_i \varepsilon_t \qquad (i = 1, \ldots, k). \end{aligned} \tag{3}$$

The model `egarch_kj` includes k volatility factors v_{it} and the normal mixture has j components ($k = 1, 2$; $j = 1, 2$).

The high-frequency models use a volatility signal extracted from five-minute intraday S&P 500 Index futures returns. Following Andersen et al. (2001), Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2002), daily realized volatility was calculated by summing

over squared intraday returns for each day t,

$$RV_t^{(\Delta)} = \sum_{j=1}^{1/\Delta} \left(f_{t-1+j\Delta} - f_{t-1+(j-1)\Delta} \right)^2, \tag{4}$$

where f_t is the log futures price and Δ is the sampling interval for the intraday data. In the application Δ corresponds to five-minute intervals. In (4) $t-1$ denotes the opening of the market on day t and t denotes the close (so intraday volatility does not include the return from market close on one day to market open on the following day).

In principle, high-frequency returns are capable of providing very precise information about the latent volatility state. In practice, there is measurement error related to, for example, market microstructure effects and non-synchronous trading, which the use of five-minute futures returns is intended to help alleviate (longer sampling intervals decrease the measurement error but at the cost of greater discretization error). Perhaps more critically, we are using the realized volatility observed on day t as a basis for forecasting day $t+1$ returns. Consistent with the literature, we also ignore the overnight return. So there is little reason to expect the realized volatility to be either an efficient or unbiased estimator for the variance of the next day's return. We address these issues in two steps. First, we apply a filter to extract estimates of the latent volatility state from the realized volatility observations. We then apply a mapping of flexible functional form from the volatility state to σ_t to compensate for any bias.

To implement the first step (filtering), we use a standard linear Gaussian state space representation for the dynamics of RV_t. Let v_t denote a latent volatility state with dynamics

$$\begin{aligned} v_t &= \mu_V + \sum_{i=1}^{k} \alpha_{it} \\ \alpha_{it} &= \phi_{it}\alpha_{i,t-1} + \sigma_{\alpha i}\eta_{it} \qquad (i = 1, \ldots, k) \end{aligned}$$

where α_{it} $(i = 1, \ldots, k)$ are the unobserved factors and η_{it} $(i = 1, \ldots, k)$ are independent

standard normals. We observe a noisy signal,

$$\log RV_t = v_t + \sigma_v \omega_t,$$

where the ω_t are *iid* standard normal and independent from η_{it} $(i = 1, \ldots, k)$. From this, an estimate of the volatility state

$$\hat{v}_t = E[v_t | RV_1, \ldots, RV_{t-1}]$$

is easily obtained using the Kalman filter. See e.g. Hamilton (1994) for details.

We also tried two alternative approaches to extracting a volatility signal from the observed RV data: an exponential weighting filter (e.g., Maheu and McCurdy, 2007) and the heterogeneous autoregressive model of Corsi (2009). The Kalman filter performed slightly better than these with respect to predictive performance, but the differences among these alternatives were small. All improved predictions substantially relative to using the unfiltered RV data directly as a proxy for the volatility state.

The model is completed by a mapping $\psi : \hat{v}_t \longrightarrow \log \sigma_t$, which we construct using flexible parametric methods. Polynomial expansions of sufficiently high degree are capable of approximating any smooth function to arbitrary accuracy on compact sets, and so are useful for this purpose. We looked at Legendre polynomials up to order three (the volatility states were first scaled and translated to mean zero and unit variance), but found no improvements beyond order two.

As in related work by Koopman and Scharth (2011), estimation takes place in two steps: first, the volatility states are extracted using the Kalman filter; then the parameters of the mapping are estimated simultaneously with the parameters of (1) conditional on the point estimates for the volatility state. In particular, note that the volatility filters do not depend upon either the mapping or (1).

We note that some efficiency is lost by using this two-step estimation procedure. But

the approach is consistent with the objective of this study, which is to demonstrate the utility of model pooling using volatility forecasts σ_t based on alternative information sets (realized volatility in this case). To the extent that the high-frequency data are much more informative about volatility than are the daily returns, the efficiency loss should be small. Additional detail illustrating the performance of the filters in practice is provided in Section 4. See Dobrev and Szerszen (2010) and Koopman and Scharth (2011) for related work. Also, there are a number of alternatives to (4) for estimating realized volatility in the literature, some of which may be more efficient than (4). See for example, Zhang et al (2005), Andersen et al (2007), Barndorff-Nielsen et al (2008) and Jacod et al (2009). We do not undertake a comprehensive study of these here.

The model `hifreq_kjp` uses a state space representation with k independent latent volatility factors, j normal components in the mixture for ε_t, and a polynomial of order p in the mapping. We consider the cases $(k = 0, 1, 2;\ j = 1, 2;\ p = 0, 1, 2)$ for a total of 18 high-frequency models. The case $k = 0$ indicates that no filtering is done (that is, $\hat{v}_t = E[v_t|RV_1, \ldots, RV_{t-1}] = RV_{t-1}$). The case $p = 0$ refers to a linear polynomial where the constant is estimated and the slope coefficient is one.

The options models have the same structure as the high-frequency models except that they substitute a measure of option-implied volatility IV_t in place of the high-frequency measure RV_t. We use the VIX index, a model-free measure of volatility implied by options prices (Britten-Jones and Neuberger, 2000). There is some measurement error involved when using the VIX index as a signal about the volatility state due to, for example, truncation and discreteness effects (Jiang and Tian, 2007). The measure is also biased due to the existence of risk-premia. Thus, similar considerations to those discussed in the context of the high-frequency models apply here as well.

The model `vix_kjp` uses k independent latent volatility factors, j normal components in the mixture for ε_t, and a polynomial of order p in the mapping. We consider the cases $(k = 0, 1, 2;\ j = 1, 2;\ p = 0, 1, 2)$ for a total of 18 options models.

Complementary to the pooling approach used in this paper, there has also been some work toward constructing unified models that combine the information from various sources (e.g., Engle and Gallo 2006; Shephard and Sheppard 2010). While we do not include such models in the analysis here, it would be straightforward to expand the pool in such directions, possibly yielding even better performance.

3 Pooling

Each model just described provides a sequence of conditional probability densities

$$p_t\left(y_t \mid Y_{t-1}^o, X_{t-1}^o, \theta_i, A_i\right) \quad (i = 1, \ldots, n = 42)$$

for the asset return y_t on day t conditional on information available at the close of trading on day $t-1$ and a parameter vector θ_i. The superscript "o" denotes the observed value (data) as distinguished from the *ex ante* random variable or argument of the density function and A_i indicates a particular model. The symbols Y_{t-1} and X_{t-1} indicate the set of asset returns and a set of covariates, respectively, on trading days $t-1, t-2, \ldots$. In the high-frequency models X_{t-1} consists of the five-minute intraday returns on days $s < t$; for the options models it consists of the VIX index on days $s < t$; and for the daily models $X_{t-1} = \emptyset$. This section uses this generic notation throughout.

Predictive densities $p_t\left(y_t; Y_{t-1}^o, X_{t-1}^o, A_i\right)$ are constructed for each model A_i and observation date t by eliminating the unknown parameter vector θ_i. This can be done in a variety of ways. One is to substitute the maximum likelihood estimate

$$\widehat{\theta}_i\left(t\right) = \arg\max_{\theta_i} \sum_{s=1}^{t-1} \log\left[p_s\left(y_s^o \mid Y_{s-1}^o, X_{s-1}^o, \theta_i, A_i\right)\right] \tag{5}$$

for θ_i, and that is the procedure used in this paper. An alternative would be to specify prior distributions $p\left(\theta_i \mid A_i\right)$ and then replace the conditional densities $p_t\left(y_t \mid Y_{t-1}^o, X_{t-1}^o, \theta_i, A_i\right)$

by the full Bayesian predictive densities.

Formal decision-making, however, requires a *single* predictive density $p_t\left(y_t; Y_{t-1}^o, X_{t-1}^o\right)$ at the end of each trading day $t-1$. Broadly speaking these contexts include any situation in which normative behavior presumes a subjective distribution for relevant unknown magnitudes, including conventional expected utility maximization. Special cases are conventional approaches to asset derivative pricing and prediction. A decision-maker could choose among the alternative predictive densities $p_t\left(y_t; Y_{t-1}^o, X_{t-1}^o, A_i\right)$ or combine them in some fashion.

3.1 Assessing the performance of predictive densities

Model choice or combination is itself a decision problem that requires a criterion. The decision-maker can use the observed values of past returns and covariates Y_{t-1}^o and X_{t-1}^o to assess the performance of any stipulated predictive density, just as an investor can use the history of returns to assess portfolio performance. This set of primitives—the history $\left(Y_{t-1}^o, X_{t-1}^o\right)$ and the predictive density function p_t—is the one typically used in the few studies that have addressed these questions (e.g., Diebold et al., 1998, p. 879). As Gneiting et al. (2007, p. 244) notes, the assessment of a predictive distribution on the basis of $p_t\left(y_t^o; Y_{t-1}^o, X_{t-1}^o\right)$ only is consistent with the prequential principle of Dawid (1984). These assessment procedures are widely known as scoring rules.

This study uses the log scoring rule

$$LS\left(Y_{t-1}^o; X_{t-1}^o, R\right) = \sum_{s=1}^{t-1} \log p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, R\right). \tag{6}$$

to assess the prediction performance of any rule R for selection or combination of predictive densities. This rule is easy to interpret, grounded in the literature, and has a significant axiomatic justification. With regard to interpretation, there is a simple relationship between

(6) and the performance of alternative rules in prediction. For alternative rules R_1 and R_2,

$$\Delta(R_1, R_2) = \exp\left\{\left[LS\left(Y_{t-1}^o; X_{t-1}^o, R_1\right) - LS\left(Y_{t-1}^o; X_{t-1}^o, R_2\right)\right] / (t-1)\right\} \tag{7}$$

is the geometric average of the ratio of probability densities assigned to the observed returns $y_1^o, \ldots, y_{t-1}^o$. This justifies the colloquial interpretation, "observed returns were $100 \cdot [\Delta(R_1, R_2) - 1]$ percent more probable under predictive density R_1 than they were under R_2."

With reference to the econometrics literature, for the specific case of Bayesian predictive densities $LS\left(Y_{t-1}^o; X_{t-1}^o, A_i\right)$ is the log predictive likelihood. In the even more specific case in which the sample begins at time $t = 1$ and sample size is T, $LS(Y_T^o; X_T^o, A_i)$ is the log marginal likelihood, which in turn is the foundation of the Bayesian approach to the model combination issue addressed in this study. (On predictive and marginal likelihoods see Geweke, 2005, Section 2.6.) Evaluation of log scores using models with maximum likelihood estimates (5), employed in this paper, is an out-of-sample criterion, and as such does not lead to complications of over-fitting.

With reference to the finance literature, the rule (6) is formally similar to a separable utility function in which the quantity of the single good consumed in period s is $p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o\right)$ and instantaneous utility is logarithmic. In the prototypical situation consumption is return on wealth and the motivating problem is optimal portfolio allocation. Higher $p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o\right)$ is better than lower just as more consumption is preferred to less.

With reference to the statistics literature, (6) is the unique proper local scoring rule, as discussed in Geweke and Amisano (2011, p 131).

3.2 Combining predictive densities

From $p_s(y_s; Y_{s-1}, X_{s-1}, A_i)$ $(s < t; i = 1, \ldots, n)$ and $\left(Y_{t-1}^o, X_{t-1}^o\right)$ the decision-maker creates $p_t\left(y_t; Y_{t-1}^o, X_{t-1}^o\right)$. We refer to this mapping as a prediction pool, motivated by the more

general descriptor opinion pool for a combination of subjective probability distributions originating with Stone (1961). There are endless ways in which the n predictive densities could be combined; see Genest et al. (1984) for a review and axiomatic approach. Restricting consideration to linear combinations leads to computations that are simple, both absolutely and in comparison with alternatives.[1] At the close of trading day $t-1$ the predictive density for the next trading day's return using a linear prediction pool is

$$p\left(y_t; X_{t-1}^o, Y_{t-1}^o, \mathbf{w}_{t-1}\right) = \sum_{i=1}^{n} w_{t-1,i} p_t\left(y_t; Y_{t-1}^o, X_{t-1}^o, A_i\right) \tag{8}$$

where $\mathbf{w}_{t-1} = (w_{t-1,1}, \ldots, w_{t-1,n})'$ is a weight vector satisfying

$$\sum_{i=1}^{n} w_{t-1,i} = 1; \;\; w_{t-1,i} \geq 0 \quad (i = 1, \ldots, n)\,. \tag{9}$$

These restrictions are sufficient to assure that (8) is a density function. Applying the log scoring rule, this linear prediction pool is scored using

$$f_{t-1}\left(\mathbf{w}_{t-1}\right) = \sum_{s=1}^{t-1} \log\left[\sum_{i=1}^{n} w_{t-1,i} p_s\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_i\right)\right] \tag{10}$$

and therefore the optimal weight vector $\mathbf{w}_{t-1}^*$ is chosen to maximize (10). The optimal weight vector is updated at the close of trading each day, reflecting the performance of the models in predicting that day's return. Geweke and Amisano (2011) shows that $f_t\left(\mathbf{w}_t\right)$ is at least weakly concave, and for $t \geq n$ $f_t\left(\mathbf{w}_t\right)$ is in general strictly concave. Maximization of f_t is therefore a regular convex programming problem and the optimal weights can be computed using conventional software.[2]

The intuition behind optimal pooling under a log scoring rule is similar to that of portfolio optimization under the constraint of no short positions. Model A_1 may have a log score that substantially exceeds that of model A_2, just as one asset may have an average return substantially higher than another. But it may also be the case that from time to time

$p_t\left(y_t^o; Y_{t-1}^o, X_{t-1}^o, A_1\right)\ p_t\left(y_t^o; Y_{t-1}^o, X_{t-1}^o, A_2\right)$ is small, much closer to zero than one, just as the asset with lower average return may from time to time substantially outperform the other. Given the concavity of the log score function, the optimal pool can (and often does) assign positive weight to both models, just as given risk aversion both assets may have positive weights in an optimal portfolio.

[Figure 1 about here]

To illustrate this intuition, suppose that there exists a data generating process D—an assumption not made to this point. Figure 1 pertains to a case in which y_t is independent and identically distributed, with the probability density function under D indicated by the solid line. Model 1 closely tracks the data generating process, except for the left lobe that is reflected in realizations about one observation in twenty. The log score of Model 2 is much lower than that of Model 1, which will be assigned negligible posterior probability in a formal Bayesian approach and be rejected in favor of Model 1 in a formal sampling-theoretic test. Yet it receives positive weight in the pool, which has a substantially higher log score than Model 1, because relative to Model 1 the pool provides a very large increase in the log predictive density when realizations from the left lobe occur.

Given further weak regularity conditions $t^{-1}\sum_{s=1}^{t}\log p\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_i\right)$ tends to an almost sure limit $L\left(A_i, D\right)$. Geweke and Amisano (2011) shows that under these conditions both the function $t^{-1}f_t\left(\mathbf{w}_t\right)$ and the sequence of optimal weight vectors $\{\mathbf{w}_t^*\}$ have well-defined almost sure pointwise limits. In general several components of the limiting weight vector are positive. An exception is the hypothetical case $D = A_i$, for which w_{ti}^* has limiting value one (Geweke and Amisano, 2011, Theorems 1 and 2). Thus in large samples several of the competing models may enter the optimal pool. This occurs because all of the models under consideration are false.

3.3 Alternatives to optimal pooling

These conditions are most explicit in Bayesian econometrics, which provides a logically complete theory of model combination. That approach explicitly conditions on one of the models being true, i.e. $D = A_i$ for some (unknown) $i = 1, \ldots, n$; Bernardo and Smith (1994, Section 6.1.2) provides an illuminating discussion of this point. Let ρ_i and π_{ti} denote the prior probability and the posterior probability of model A_i conditional on the first t observations, respectively. For any pair of models A_i and A_j,

$$\log\left(\frac{\pi_{ti}}{\pi_{tj}}\right) = \log\left(\frac{\rho_i}{\rho_j}\right) + \sum_{s=1}^{t} \log p\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_i\right) - \sum_{s=1}^{t} \log p\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_j\right).$$

Granted the existence of a data generating process D, $t^{-1}\log\left(\pi_{ti}/\pi_{tj}\right) \xrightarrow{a.s.} L\left(A_i, D\right) - L\left(A_j, D\right)$. Since this limit is generally positive or negative, $\log\left(\pi_{ti}/\pi_{tj}\right)$ tends either to $+\infty$ or $-\infty$ as $t \to \infty$. In general there is one model A_i for which $\log\left(\pi_{ti}/\pi_{tj}\right) \to +\infty$ $(j \neq i)$. In the Bayesian model averaging pool the weight on the predictive densities of model A_i tends to 1 and the weights on the predictive densities of all other models tend to 0 as $t \to \infty$. This phenomonen was noted by Diebold (1991).

Bayesian model averaging and optimal model pooling lead to very different choices of a single predictive density $p\left(y_t; Y_{t-1}^o, X_{t-1}^o\right)$. The conclusions are different because the assumptions are different. Bayesian model averaging conditions on the data generating process being one of the models under consideration. Granted this condition, as evidence accumulates that a particular model A_i is superior to all the others, one is driven to the conclusion that $A_i = D$ and predictions should be based on that model alone. Model pooling does not assume the existence of a data generating process of any kind, although this assumption is convenient for establishing asymptotic properties of prediction pools. If the stronger conditions of Bayesian model averaging are in fact correct then model pooling leads to the same result asymptotically.

Bayesian model averaging is preferable for prediction given a strong prior belief that one

of the models under consideration is the process responsible for the data. Model pooling is preferable in the absence of this belief and the appropriateness of the log scoring rule for the prediction problem at hand. We find the conditions for Bayesian model averaging excessively strong for predicting daily financial returns, but others may not. Regardless, a finding that model pooling systematically produces higher log predictive scores than does Bayesian model averaging is evidence against the proposition that one of the models corresponds to the data generating process. That turns out to be the case here.

4 Results

The application uses S&P 500 Index (SPX) log returns from January 2, 1990 through March 31, 2010. Models based on option-implied volatility use the VIX index. SPX and VIX data were obtained directly from the Chicago Board Options Exchange (CBOE). Following Andersen, Bollerslev, Diebold and Vega (2007), the high-frequency models use a volatility signal extracted from five-minute intraday S&P 500 Index futures returns (obtained from TickData.com). We also experimented with estimating the high-frequency models using a volatility signal extracted from five-minute intraday returns for the S&P 500 cash index itself (rather than index futures). But the five-minute returns for the cash index exhibit strong positive serial correlation due to the fact that there may be few or no trades reported in some of the stocks in the index in a given 5-minute interval. This non-synchronous trading issue is mitigated by using futures prices instead, resulting in better volatility estimates and improved predictive performance.

Since the VIX begins with the first trading day of 1990, estimation samples for all of our models begin with t corresponding to the second trading day of 1990. For each model A_i we evaluate predictive densities $p\left(y_t^o; Y_{t-1}^o, X_{t-1}^o, A_i\right)$ recursively, beginning with $t = 1$ corresponding to the first trading day of 1992 and ending with $t = T = 4596$ corresponding to March 31, 2010. This requires re-estimation of each model for each t as $\mathbf{Y}_{t-1}^o$ expands.

Since there are 4596 days in the recursion and 42 models, the result is a 4596×42 matrix $\mathbf{P}$ of predictive densities. These computations are relatively time consuming.[3] All of our findings derive from $\mathbf{P}$.

4.1 Discussion of volatility filters

[Figures 2 and 3 about here.]

We begin by examining the performance of several of the volatility filters described in Section 2. Figure 2 shows estimated volatility states, $\hat{v}_t = E[v_t|RV_1, \ldots, RV_{t-1}]$, for the high-frequency models with one- and two-factor Kalman filters. Figure 3 is analogous for the options models. Recall that these filters do not depend on the other model features, either the mapping ψ from states to σ_t or (1). For reference, we also show the volatility states corresponding to unfiltered data ($\hat{v}_t = \log RV_{t-1}$ for the high-frequency models or $\hat{v}_t = \log IV_{t-1}$ for the options models). For clarity, we show only the last six months of the sample.

Examination of Figure 2 suggests that the Kalman filter is effective in removing noise from the observed values of RV. Most of the work is done by the first factor. Including the second factor has only a small effect at the one-day horizon considered here (the impact may be greater on forecasts at longer horizons since the two-factor models generate much more persistence in volatility than do the single-factor models). Supporting evidence is provided by the large improvements in log score reported in Section 4.2 for models using the single-factor Kalman filter and smaller additional improvements for models that use the two-factor filter.

For the options models, on the other hand, the Kalman filter does little (Figure 3). This is also supported by the results for log score reported in Section 4.2 (log scores for models using the single-factor Kalman filters are in most cases essentially identical to those using unfiltered IV; including a second volatility factor reduces log score slightly, typical of over-fitting in a predictive setting).

[Figure 4 about here.]

Figure 4 shows the mappings $\psi : \hat{v}_t \longrightarrow \log \sigma_t$ corresponding to several high-frequency and options models. Mappings corresponding to $p \in \{0, 1, 2\}$ conditional on the full sample are shown. Recall that $p = 0$ corresponds to a linear mapping with the linear coefficicent fixed at one, $p = 1$ to an unconstrained linear mapping, and $p = 2$ to a quadratic mapping. The 45 degree line (corresponding to $\log \sigma_t = \hat{v}_t$) is shown for reference. For the high-frequency models, there is little difference among these except at the extremes of the range of observed data, suggesting that the simplest formulation is largely sufficient. This conclusion is supported by the log scores reported in Section 4.2, which show only small differences corresponding to alternative formulations. For the options models, the differences are larger. The figure suggests that the constraint on the linear coefficient implied by $p = 0$ is not supported by the data. The quadratic mapping is close to the (unconstrained) linear mapping through the region where most of the data occurs, but diverges from it at very high values of IV. The log scores reported in Section 4.2 show substantial improvement when going from $p = 0$ to $p = 1$, but log scores are worse for $p = 2$, indicative of over-fitting. We also tested cubic mappings for both high-frequency and options models but there was never any benefit to these and we do not report these results.

Note that the mappings for the high-frequency models are all above the 45 degree line, implying that RV is a downward biased estimate of $\log \sigma_t$. As discussed in Section 2, possible causes for this bias include market micro-structure effects, non-synchronous trading issues, and the fact that RV does not incorporate overnight returns. For the options models the mappings are predominately below the 45 degree line, indicating that IV is on average an upward biased estimate of $\log \sigma_t$ (reflecting the existence of a volatility risk premium). The differing biases associated with the IV and RV measures of volatility are also evident in Figures 2 and 3 (IV tends to be higher than RV). The different mappings from implied volatility state to σ_t attempt to correct for this discrepancy.

[Figure 5 about here.]

Figure 5 shows values of σ_t implied by several models. The various volatility measures track generally close to each other, but there are occasional persistent divergences. It is largely these differences in implied values of σ_t that underlie the variation in performance for predicting returns amongst the models.

4.2 Model performance and comparison

Table 1 provides the (full sample) log predictive score (6) of each model, $LS\left(Y_T^o; X_T^o, A_i\right)$ $(i = 1, \ldots, n)$. For legibility we subtract the log predictive score of the `hifreq_010` model, which is 14,900.39, from the log scores reported here and throughout Section 4. Differences in log scores, not their levels, matter. From (7), the difference $\Delta\left(A_i, A_j\right) = \exp\left\{\left[LS\left(Y_T^o; X_T^o, A_i\right) - LS\left(Y_T^o; X_T^o, A_j\right)\right]/T\right\}$ corresponds to a geometric average proportional difference in predictive densities. For example in the case of `hifreq_222` and `hifreq_010` this difference is $\exp\left(206.89/4596\right) = 1.046$. That is, the predictive densities from model `hifreq_222` render observed events on average almost 5% more probable than do the predictive densities from model `hifreq_010`. More generally, a difference of 45.73 in log scores corresponds to a 1% increment in probability, a difference of 4.59 to a 0.1% increment.

In interpreting the results, it is essential to recall that the log predictive score is an out-of-sample criterion. Unlike in-sample criteria, out-of-sample criteria inherently penalize overfitting. If model A_i is nested in model A_j, the predictive likelihood of model A_i can exceed that of model A_j; in contrast, the maximized log-likelihood (an in-sample criterion) can never be higher for the nested model. In Table 1 notice that the `vix_121` model is nested in the `vix_222` model and has the higher log score; similarly for `hifreq_120` and `hifreq_121`.

As noted in Section 3, had our method of inference been formally Bayesian, then the log scores would coincide with marginalized likelihoods in which the prior distribution for each model includes the 1990-1991 data as a training sample. That is not the case here, but differences in log scores can be regarded as of the same order of magnitude as log ratios of posterior probabilities. For example, given equal prior probablilities for the models, the

posterior probability odds ratio in favor of `sv_2` over `sv_1` is on the order of 10^9.

[Table 1 about here.]

This interpretation reveals the high return to the various elaborations on the daily, high-frequency and options models detailed in Section 2. The roughly 20-point improvement for the stochastic volatility model, resulting entirely from using a mixture of normals rather than Gaussian distribution for ε_t, has just been noted. Returns for other model classes are higher. Among the EGARCH models, `egarch_22` improves over the conventional model, `egarch_11`, by over 100 points with the introduction of a second volatility factor and use of a mixture distribution for the shocks. The improvement is most dramatic for the high-frequency models, where the increase of over 200 points in log score relative to the simplest model is due primarily to the incorporation of a filtration ($k > 0$) that allows current latent volatility to depend flexibly on lagged realized volatilities and secondarily to the use of a mixture distribution for the return shocks ($j = 2$). For the options models the elaborations described in Section 2 lead to an increase of over 80 points in log score, accounted for primarily by the mixture of normals distribution for conditional returns and secondarily by the incorporation of additional flexibility in the link between IV_t and σ_t ($p = 1$ versus $p = 0$).

4.3 Conventional predictive density combination

Arguably the simplest rule for density combinations is the equally-weighted pool A^*, which has $w_{i,t-1} = n^{-1}$ $(t = 1, \ldots, T; i = 1, \ldots, n)$ and log score $LS\left(Y_T^o; X_T^o, A^*\right)$. From Jensen's inequality $LS\left(Y_T^o; X_T^o, A^*\right)$ must exceed the mean log predictive score in Table 1. Indeed it can exceed the maximum of the log predictive scores, and that is what happens here: $LS\left(Y_T^o; X_T^o, A^*\right) = 231.03$.

A modest elaboration on this procedure is first to distribute weight equally on each group of models and then equally across models within each group. Thus in this application

each group has weight 1/3, so that each daily model has weight 1/18 and each of the high-frequency and options models has weight 1/54. The log score of the resulting pool is 233.86.

Equally-weighted pools provide useful benchmarks for comparisons with alternative predictive density rules. The idea is similar to the use of the market portfolio or $1/n$ rules as benchmarks for portfolio performance. Many stock pickers believe that they can reliably beat the market. Far fewer succeed. The analogy holds for model selection as well.

The best performing individual model over the full sample period is `hifreq_222`. An econometrician using a conventional approach to select a single "best" model would place all weight on this model to the exclusion of alternatives. But even the simplest equally-weighted pool beats this model by over 24 points in log score.

The reality for the model picker is even worse than this. Here, we have assumed a prescient model picker who is able to choose the individual model that performs best over the entire sample. In practice, the model picker must choose the best model in real time using available information. The real-time model picker underperforms the equally-weighted pool by nearly 40 points in log score.

Bayesian model averaging (BMA) is often put forward as an appealing approach to model combination. It is instructive to consider the idea of constructing real-time pools using BMA in order to examine the implications for choices amongst the 42 individual models and for contrasting these implications with optimal pooling subsequently.

Identifying $p\left(y_t^o; Y_{t-1}^o, X_{t-1}^o, A_i\right)$ with the Bayesian predictive likelihood, the analogue of marginal likelihood for model A_i based on the sample from periods 1 through t is

$$ML_{it} = \prod_{s=1}^{t} p\left(y_s^o; Y_{s-1}^o, X_{s-1}^o, A_i\right) = \exp\left[LS\left(Y_t^o; X_t^o, A_i\right)\right].$$

Given equal model prior probabilities, the posterior probability of model i based on this sample is $\omega_{it} = ML_{it} / \sum_{j=1}^{n} ML_{jt}$. Under the Bayesian model averaging paradigm, the

predictive density for y_{t+1} is

$$p\left(y_{t+1}; Y_t^o, X_t^o, B^*\right) = \sum_{i=1}^{n} \omega_{it} p\left(y_{t+1}; Y_t^o, X_t^o, A_i\right). \tag{11}$$

The procedure just described constitutes a valid prediction model, which we denote B^* in (11). Its log predictive score $LS\left(Y_T^o; X_T^o, B^*\right)$ can be evaluated directly using the 4596×42 matrix $\mathbf{P}$ of predictive likelihoods described at the beginning of this section.

[Figure 6 about here.]

Two features of this model averaging exercise are important for this study. First, consider the weights ω_{it}. Rather than report weights for all of the models individually, at each time period t we sum the weights within each of the three groups of models (daily, high-frequency and options). These are displayed in Figure 6. In the early part of the sample, the preponderance of the weight is on the options model group, with occasional reversals in favor of the daily models. Then beginning around 2001 the high-frequency models dominate, except for late 2007 through late 2008 when the daily models again take most of the weight. Toward the very end of the sample (from late 2009) weight is about evenly split between the high-frequency and daily models. Almost no weight is placed on any of the options models after mid-2000. The tendency of Bayesian model averaging to concentrate weight on a single model was noted nearly two decades ago in Diebold (1991). The vacillation between near-certainties exhibited in Figure 6 implied by a procedure that starts with the premise that one of the models corresponds to the data generating process challenges the credibility of the premise.

Second, consider the log scores. The log score of the Bayesian model averaging prediction rule is $LS\left(Y_T^o; X_T^o, B^*\right) = 203.73$. Though this is slightly better than the modeler who places all weight on a single model in real time, it still falls well short of the benchmark equally-weighted pool (by over 25 points in log score). Thus the performance of this procedure motivated by Bayesian model averaging is poor just as its premise is not credible.

4.4 Optimal pooling

The optimal pooling procedure implemented here reconstructs what an econometrician could have accomplished in real time. For each date t beginning with $t = 1$, which indicates the first trading day of 1992, and ending with $t = 4596$ (March 31, 2010) suppose that the econometrician has at her disposal predictive densities $p_s(y_s; Y^o_{s-1}, X^0_{s-1}, A_i)$ $(s = 1, \ldots, t; i = 1, \ldots, n)$ and has evaluated these densities using realized returns, $p_s(y^o_s; Y^o_{s-1}, X^0_{s-1}, A_i)$. Thus, on day t, the optimizer is using the first t rows of the 4596×42 matrix $\mathbf{P}$. Using this information, she finds the optimal pooling weights $\mathbf{w}^*_t = \arg\max_{\mathbf{w}_t} f_t(\mathbf{w}_t)$ where $f_t(\mathbf{w}_t)$ is defined in (10).

[Figure 7 about here.]

Figure 7 displays the optimal pool weights w^*_{it} in the same way that Figure 6 did for the Bayesian model averaging weights. Initially the optimal pool consists entirely of daily models. High-frequency models enter the optimal pool midway through the first year and options models enter shortly thereafter. The gradual entry of models at the start of the exercise is characteristic of optimal prediction pools: notice from the calculus of optimization of a concave function on the unit simplex in (8)-(9) that at most t models will have positive weight in an optimal pool when $t < n$. As the number of predictions over which the optimal pool weights are evaluated continues to increase, optimal weights stabilize. From midway through the exercise (2001) forward the distribution of weights across the groups of models does not change substantially.

At the end of the exercise, which is the close of trading on March 31, 2010, the total weight on the group of daily models is 0.426, all arising from the `egarch_22` model. The total weight on the high-frequency models is 0.406, comprised of the sum of the weights on `hifreq_010` (0.088), `hifreq_020` (0.057), `hifreq_110` (0.035), `hifreq_112` (0.080), and `hifreq_122` (0.145). The options models garner the remaining weight of 0.168, all allocated to `vix_111`. Weights for the other 35 models are all exactly zero.

Consulting Table 1, note that among the daily and options models, the variants with the highest log score get all the weight. Among the high-frequency models, in contrast, weight is distributed across five different models. The optimal pool puts no weight on `hifreq_222`, although it is the best-performing individual model. The worst performing individual model, `hifreq_010`, has positive weight.

Whether or not a model enters the pool with positive weight depends on its record in providing a higher density to observed returns when other models with positive weights provide lower densities. These conditions are analogous to those that prevail when an asset enters a portfolio under a constraint of no short positions, and arise for essentially the same reason. The optimal pool places a premium on diversity of models, even if some of those included have relatively low scores. For example, the total weight on models which include only a single mixture component is 0.371, although adding an additional mixture component substantially improves the individual models in every case.

Having computed the optimal weight vector $\mathbf{w}_t^*$ at the end of trading day t, based on rows 1 through t of $\mathbf{P} = [p_{ti}]$, our hypothetical econometrician uses the optimal pool as the predictive density for y_{t+1}. Evaluating this density at the realized return y_{t+1}^o provides the log score

$$\sum_{t=0}^{T-1} \log \left[\sum_{i=1}^{n} w_{it}^* p\left(y_{t+1}^o; Y_t^o, X_t^o\right) \right] = \sum_{t=1}^{T} \log \left(\sum_{i=1}^{n} w_{i,t-1}^* p_{ti} \right), \tag{12}$$

which may be compared directly with the entries in Table 1. The log score of the optimal pool is 238.38, about 31 points higher than the best of the constituent models, `hifreq_222`. The improvement is even greater relative to either the BMA pooling rule or the econometrician forced to place all weight on a single model using real-time information. It also exceeds the two equally-weighted benchmarks described in Section 4.3.

[Figure 8 about here.]

Figure 8 shows log scores relative to the equally-weighted pool at each date t in the sample period for the optimal pool, BMA pool using real-time weights, and the pool comprised of

the single model chosen in real-time by the model picker. Whereas the conventional model averager and model picker both substantially underperform the equally-weighted benchmark, the optimal pool outperforms it.

The sums of model weights across groups exhibited in Figure 7 provide one indication of the contribution of each group to the optimal pool. An indication more directly related to performance can be constructed as follows. First evaluate the real-time log score (12) at the end of each period t, yielding the sequence of real-time log scores

$$\lambda_t = \sum_{s=0}^{t-1} \log \left[\sum_{i=1}^{n} w_{is}^* p\left(y_{s+1}^o; X_s^o, Y_s^o\right) \right] = \sum_{s=1}^{t} \log \left(\sum_{i=1}^{n} w_{i,s-1}^* p_{si} \right) \quad (t = 1, \ldots, T) .$$

Now repeat the optimization exercise, but omitting all of the daily models, and denote the resulting sequence of log scores $\left\{\lambda_t^{(1)}\right\}$. Because of the real-time nature of the exercise it is not necessarily the case that $\lambda_t^{(1)} \leq \lambda_t$, and both prior considerations and the weights displayed in Figure 7 suggest that this condition is more likely to be violated for smaller than for larger t. We refer to $\lambda_t - \lambda_t^{(1)}$ as the value of the daily model group at time t. Similarly form the sequence of values $\left\{\lambda_t - \lambda_t^{(2)}\right\}$ for the group of high-frequency models and $\left\{\lambda_t - \lambda_t^{(3)}\right\}$ for the group of options models. Unlike sums of weights within groups, group values will tend to drift with time. For any group with a limiting positive sum of weights, the drift will be upward.

[Figure 9 about here.]

Figure 9 shows the group values constructed in this way. The value of the options models is always small and toward the middle part of the sample is even negative. The high-frequency models also have low and sometimes negative value through the early part of the sample, but their value increases dramatically from 2000 through 2004. The value of the daily models increases gradually from about 2000 through the end of the sample period, with a big jump in early 2007. At the end of the sample period (March 31, 2010) the value of the daily model group is 13.47, the high-frequency model group 18.47, and the options

group 0.53.

5 Conclusion

This study took up the practical problem of constructing predictive densities for S&P 500 returns from a collection of models, all of which are false. The constituents of the collection were chosen with respect to alternative information sets for predictions of future volatility: daily returns, observed intraday volatility, and the VIX index obtained from options prices. The metric of evaluation was the log scoring rule, equivalent to the geometric average probability assigned to observed returns. This and all comparisons made in the study are strictly out of sample, arising from real-time procedures that could have been employed in prediction at the start of each trading day from January 2, 1992, through March 31, 2010.

Beginning with conventional base models within each of the three groups, we took several steps to improve predictions: replacing conditional Gaussian distributions with normal mixture distributions provided predictive distributions with more credible shapes; including multiple volatility factors provided increased flexibility in how the history of realized returns impacted estimates of the latent volatility state; and in the case of the high frequency and options models we used a flexible mapping from the extracted volatility state to the scaling factor that determines the variance of daily returns. This led to two stochastic volatility (SV) models, four EGARCH models, 18 high frequency models and 18 options prices, for a total of 42 models.

Quantitatively this was the most important step in improving predictive densities for the S&P 500 return series from 1992 through the first quarter of 2010, as indicated in Table 2. The *Improved model* column compares the base model in each group (e.g. `hifreq_010`) with the best model in each group (e.g. `hifreq_222`) using the entries from Table 1 and the metric shown in (7). As discussed in Section 4.2 differences across model groups arise more from disparity among base models than among the best models in each group.

[Table 2 about here.]

Next we considered pools of all 42 models. The simple step of forming an equally-weighted pool of models led to the improvements in the *Equal weight pool* column of Table 2. Since the pool is the same for all model groups, differences across model groups in this column are due entirely to differences in the log predictive scores of the best model in each group. If it were not the case that all models are false—that is, some one of the 42 models in our collection corresponded to the data generating process for returns—then the expected incremental change in this column would be negative for the group containing the true model. That is far from the case. The optimal pool provides further increases in prediction probability.

Conventional econometric model combination procedures, most highly developed in the Bayesian literature, work from the condition that one of the models is true. As an alternative to optimal pooling we examined Bayesian model averaging (BMA). Whereas optimal pools lead to stable positive weights on all three groups of models, BMA weights tend to eliminate all models but one. Furthermore, the model so identified as being almost certainly true changes from time to time over the sample period. The log score of the BMA pool was lower even than that of the simple equally-weighted pool. Prediction probabilities were on average 0.76% lower for the BMA pool than for the optimal pool. The poor performance of BMA complements the incredibility of the assumption that truth resides somewhere in the collection of models.

All dimensions of the study bear out the importance of the fact that no matter what the collection of models, they are all false. Therefore improved models exist, and in this study improvement of individual models yielded the greatest returns. But even with a set of improved models, the fact that all still remain false indicates a further improvement from model pooling (Geweke and Amisano, 2011, Theorems 1 and 2). That potential was borne out in this study. This latter improvement significantly recasts model comparison from a horse race in which there is typically little role for any but the winning model to a more cooperative situation in which many models have relative strengths and weaknesses leading

to important roles for several models in improving predictive performance. In this setting an optimal pool bears strong resemblance to optimal portfolio allocation with a restriction of no short positions and the familiar gains from diversification in that setting.

Our study addressed one-step-ahead predictions of a single return, the S&P 500 index return, which in turn is the most thoroughly addressed prediction problem in financial econometrics. In contrast the most important prediction problems involve multiple returns and prediction horizons of several steps. The fundamental principles in this work, log scoring and optimal pooling, apply directly to these extensions. The case of multiple returns is straightforward, e.g. O'Doherty et al. (2010). Moreover for multivariate prediction there are compelling axiomatic arguments requiring pools to be linear (McConway, 1981) as they were in this study. Predicting several steps into the future is more demanding to the extent that covariates (in thus study, X_{t-1}, the indicators of volatility in the high frequency and options models) must also be predicted. In econometric terms these covariates are then no longer exogenous but instead must themselves be modelled. There are no fundamental difficulties, here, just the significant work of creating and improving credible models. We plan to address these issues in future research.

References

Andersen, T.G., T. Bollerslev, F.X. Diebold, and P. Ebens. 2001. The distribution of realized stock return volatility. *Journal of Financial Economics* 61: 43-76.

Andersen, T.G., T. Bollerslev, F.X. Diebold, and P. Labys. 2003. Modeling and forecasting realized volatility. *Econometrica* 71: 579-625.

Andersen T.G., T. Bollerslev, and F.X. Diebold. 2007. Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. *Review of Economics and Statistics* 89: 701-720.

Andersen, T.G., T. Bollerslev, F.X. Diebold, and C. Vega. 2007. Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics* 73: 251-277.

Barndorff-Nielsen, O.E. and N. Shephard. 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B* 64: 253-280.

Barndorff-Nielsen, O.E., P.R. Hansen, A. Lunde, and N. Shephard. 2008. Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica* 76: 1481-1536.

Bernardo, J.M. and A.F.M. Smith. 1994. Bayesian Theory. New York: Wiley.

Bolleslev, T. and H. Mikkelsen. 1996. Modelling and pricing long-memory in stock market volatility. *Journal of Econometrics* 73: 151-184.

Brier, G.W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78: 1-3.

Britten-Jones, M. and A. Neuberger. 2000. Option prices, implied price processes, and stochastic volatility. *Journal of Finance* 55: 839-866.

Corsi, F. 2009. A simple long-memory model of realized volatility. *Journal of Financial Econometrics* 7: 174-196.

Dawid, A.P. 1984. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A* 147: 278-292.

Diebold, F.X. 1991. On Bayesian forecast combination procedures. In A. Westlund and P. Hackl (eds.), *Economic Structural Change: Analysis and Forecasting*, Chapter 15, 225-232. New York: Springer-Verlag.

Diebold, F.X., T.A. Gunter, and A.S. Tay. 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39: 863-883.

Dobrev, D.P. and P.J. Szerszen. 2010. The information content of high-frequency data for estimating equity return models and forecasting risk. Working paper.

Durham, G. 2007. SV mixture models with application to S&P 500 index returns. *Journal of Financial Economics* 85: 822-856.

Engle, R.F. and G.M. Gallo. 2006. A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics* 131: 3-27.

Genest, C., S. Weerahandi, and J.V. Zidek. 1984. Aggregating opinions through logarithmic pooling. *Theory and Decision* 17: 61-70.

Geweke, J. 2005. Contemporary Bayesian Econometrics and Statistics. New York: Wiley.

Geweke, J. and G. Amisano. 2011. Optimal prediction pools. *Journal of Econometrics*, forthcoming.

Gneiting, T., F. Balabdaoui, and A.E. Raftery. 2007. Probability forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B* 69: 243-268.

Hamilton, J. 1994. Time Series Analysis. Princeton, NJ: Princeton University Press.

Jacod, J., Y. Li, P.A. Mykland, M. Podolskij, and M. Vetter. 2009. Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Processes and Their Applications* 119: 2249-2276.

Jiang, G.J. and Y.S. Tian. 2005. The model-free implied volatility and its information content. *Review of Financial Studies* 18: 1305-1342.

Koopman, S.J. and M. Scharth. 2011. The analysis of stochastic volatility in the presence

of daily realised measures. *Journal of Financial Econometrics*, forthcoming.

Maheu, J.M. and T.H. McCurdy. 2007. Components of market risk and return. *Journal of Financial Econometrics* 5: 560-590.

McConway, K.J. 1981. Marginalization and linear opinion pools. *Journal of the American Statistical Association* 76: 410-414.

McLachlan, G. and D. Peel. 2000. Finite Mixture Models. New York: Wiley.

O'Doherty, M.S., N.E. Savin, and A. Tiwari. 2010. Modeling the Cross Section of Stock Returns: A Model Pooling Approach. *Journal of Financial and Quantitative Analysis*, forthcoming.

Richard, J.F. and R. Liesenfeld. 2006. Classical and Bayesian analysis of univariate and multivariate stochastic volatility models. *Econometric Reviews* 25: 335-360.

Shephard N. and K. Sheppard. 2010. Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics* 25: 197-231.

Stone, M. 1961. The opinion pool. *Annals of Mathematical Statistics* 32: 1339-1342.

Zhang, L., P.A. Mykland, and Y. Aït-Sahalia. 2005. A tale of two time-scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100: 1394-1411.

Notes

[1]When prediction addresses vector y_t rather than scalar as is the case here, only linear combinations of predictive densities satisfy some basic conditions of internal consistency, as first shown by McConway (1981).

[2]Results reported in Section 4 all use the Matlab function `fmincon`.

[3]The stochastic volatility models required the most time, about 12 CPU days for the one-factor model and 4 CPU weeks for the 2-factor model. Models with two-factor Kalman filters took about 5 CPU hours. The other models required about 15 minutes on average. In each case the time stated is the total over all 4596 samples.

Daily models

sv_1	sv_2	egarch_11	egarch_12	egarch_21	egarch_22
159.40	180.77	98.93	169.66	139.26	**206.54**

High frequency models (hifreq_kjp)

	$j=1$			$j=2$		
	$p=0$	$p=1$	$p=2$	$p=0$	$p=1$	$p=2$
$k=0$	0.00	21.57	35.07	100.00	102.52	108.65
$k=1$	145.13	137.47	142.26	194.36	192.31	196.53
$k=2$	155.21	148.33	154.21	204.15	203.01	**206.89**

Options models (vix_kjp)

	$j=1$			$j=2$		
	$p=0$	$p=1$	$p=2$	$p=0$	$p=1$	$p=2$
$k=0$	99.36	122.69	118.34	154.37	181.16	180.57
$k=1$	92.91	122.33	118.06	149.47	**181.24**	180.73
$k=2$	89.11	117.43	112.91	146.93	177.72	176.94

Table 1: Log scores of models relative to hifreq_010. Boldface indicates the highest log score in each of the three groups of model. See Section 2 for complete model definitions.

Model group	Improved model	Equal weight pool	Optimal pool	Total
Daily (SV)	0.466	1.100	0.160	1.726
Daily (EGARCH)	2.369	0.534	0.160	3.063
High frequency	4.604	0.527	0.160	5.291
Options prices	1.798	1.089	0.160	3.047

Table 2: Improvements in the geometric mean average probability assigned to observed returns. Incremental percentage changes in prediction probability moving from left to right in each row are reported

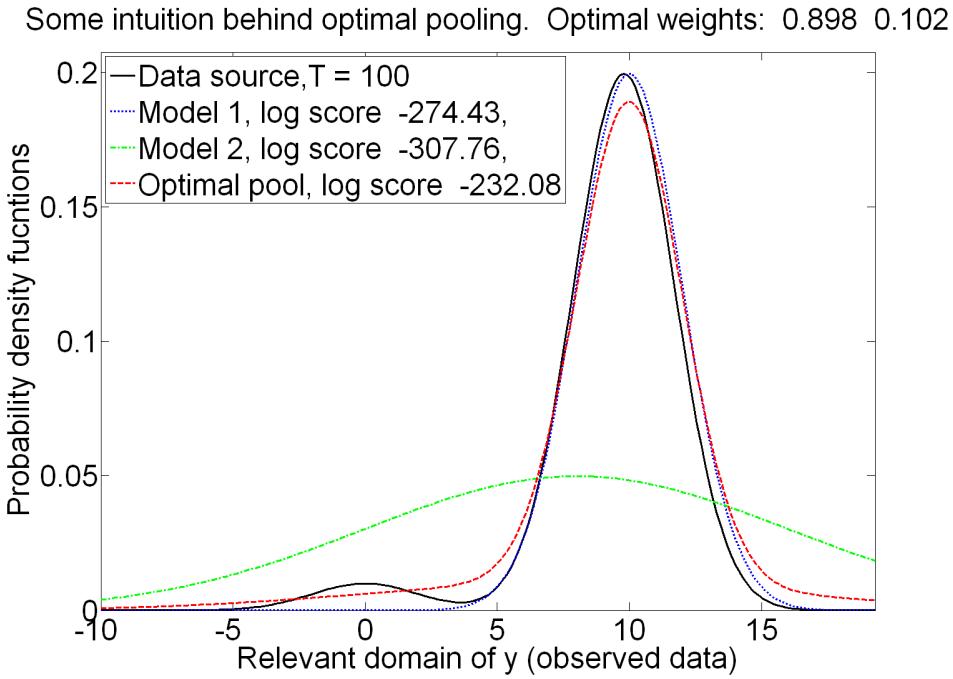

Figure 1: Constructed example illustrating optimal pooling.

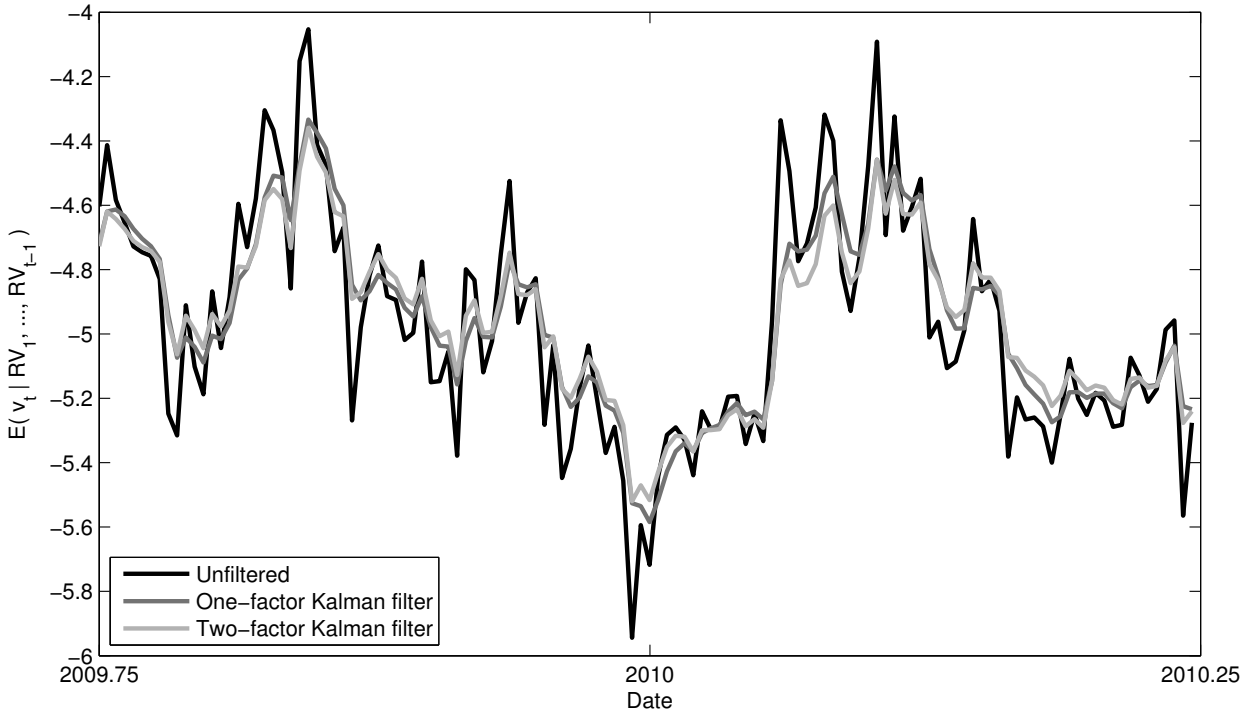

Figure 2: Filtered volatility states, high frequency models.

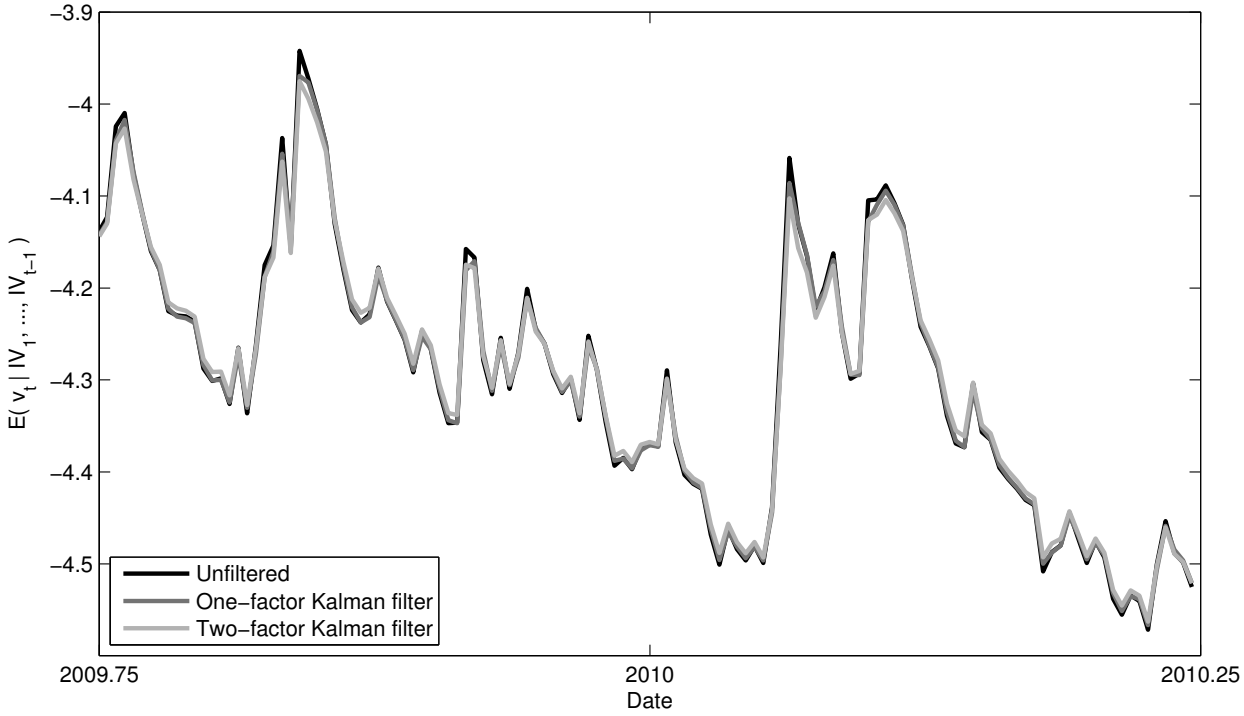

Figure 3: Filtered volatility states, options models.

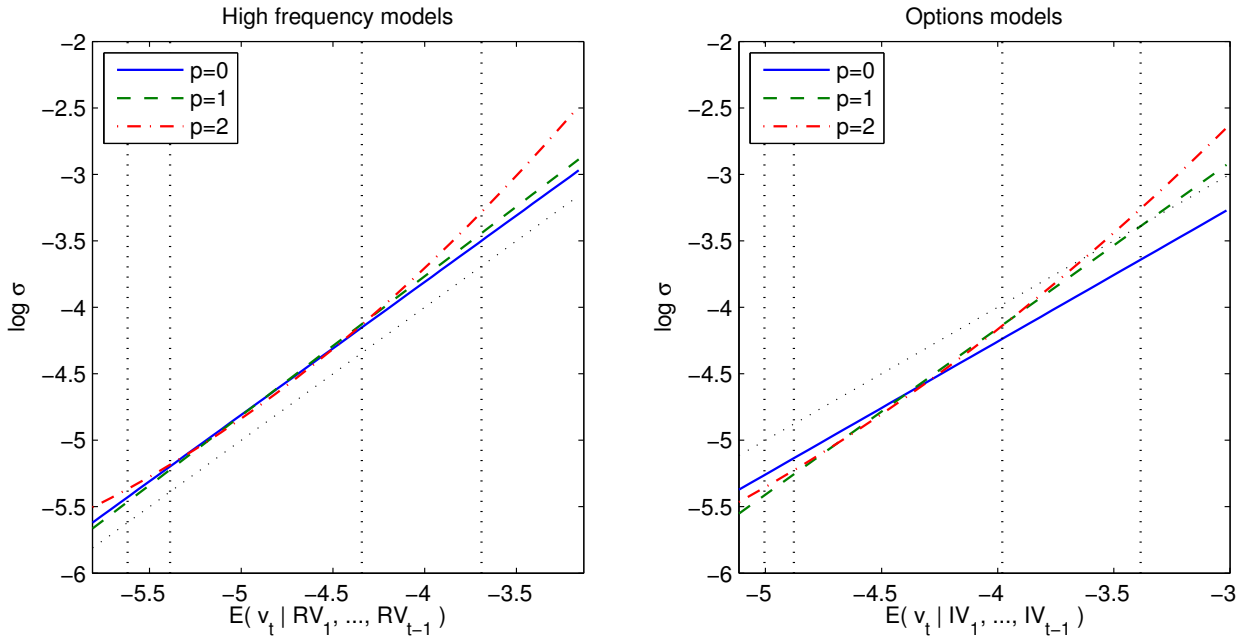

Figure 4: Mappings from volatility states to σ. All mappings are conditional on the full sample. The left panel shows mappings for `hifreq_220`, `hifreq_221` and `hifreq_222`. The right panel shows mappings for `vix_120`, `vix_121` and `vix_122`. Dotted lines indicate the 45 degree line ($\log \sigma = \hat{v}$), and the 0.01, 0.10, 0.90 and 0.99 quantiles of the data.

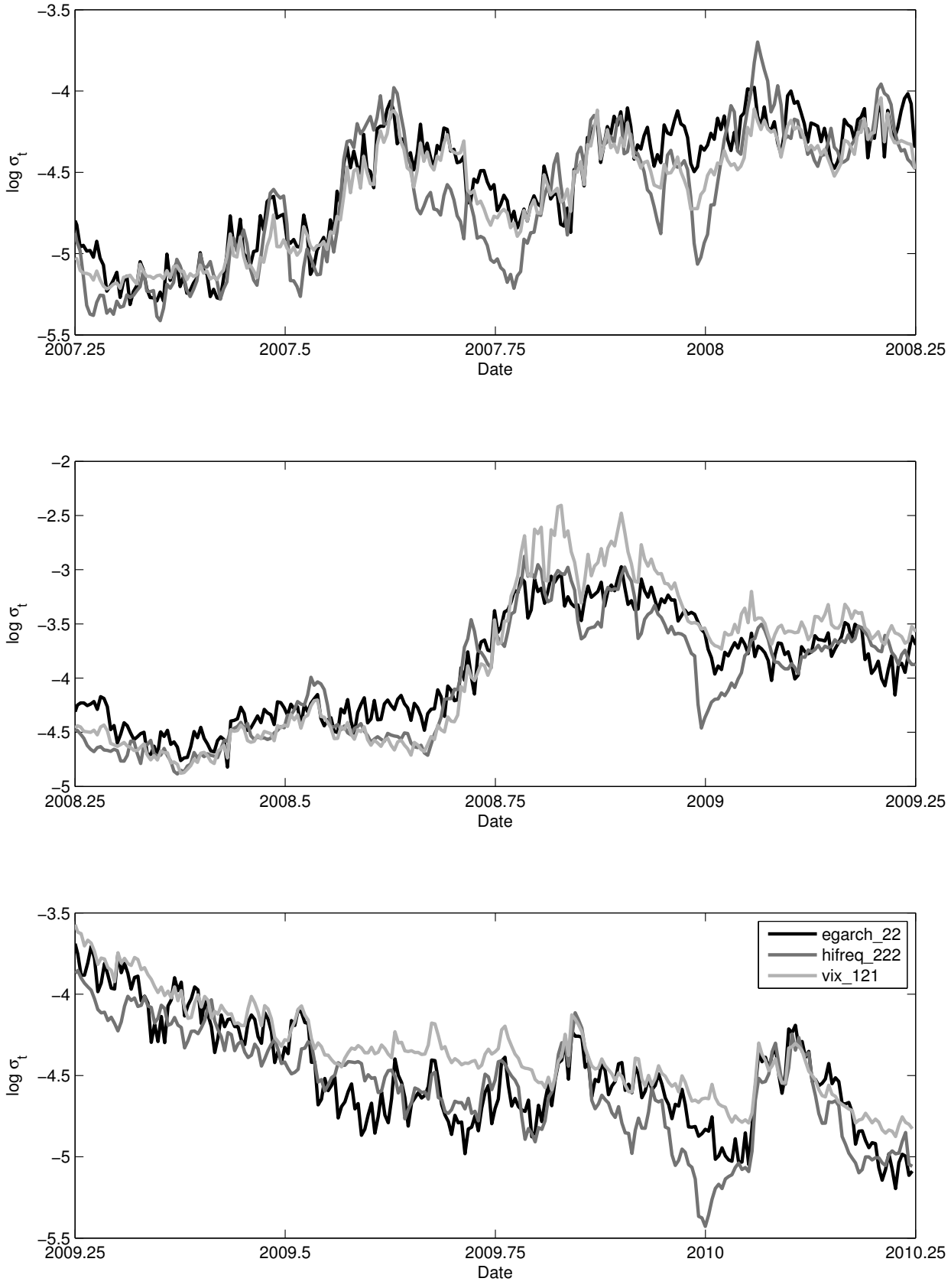

Figure 5: Implied values of σ_t corresponding to `egarch_22`, `hifreq_222` and `vix_121` models.

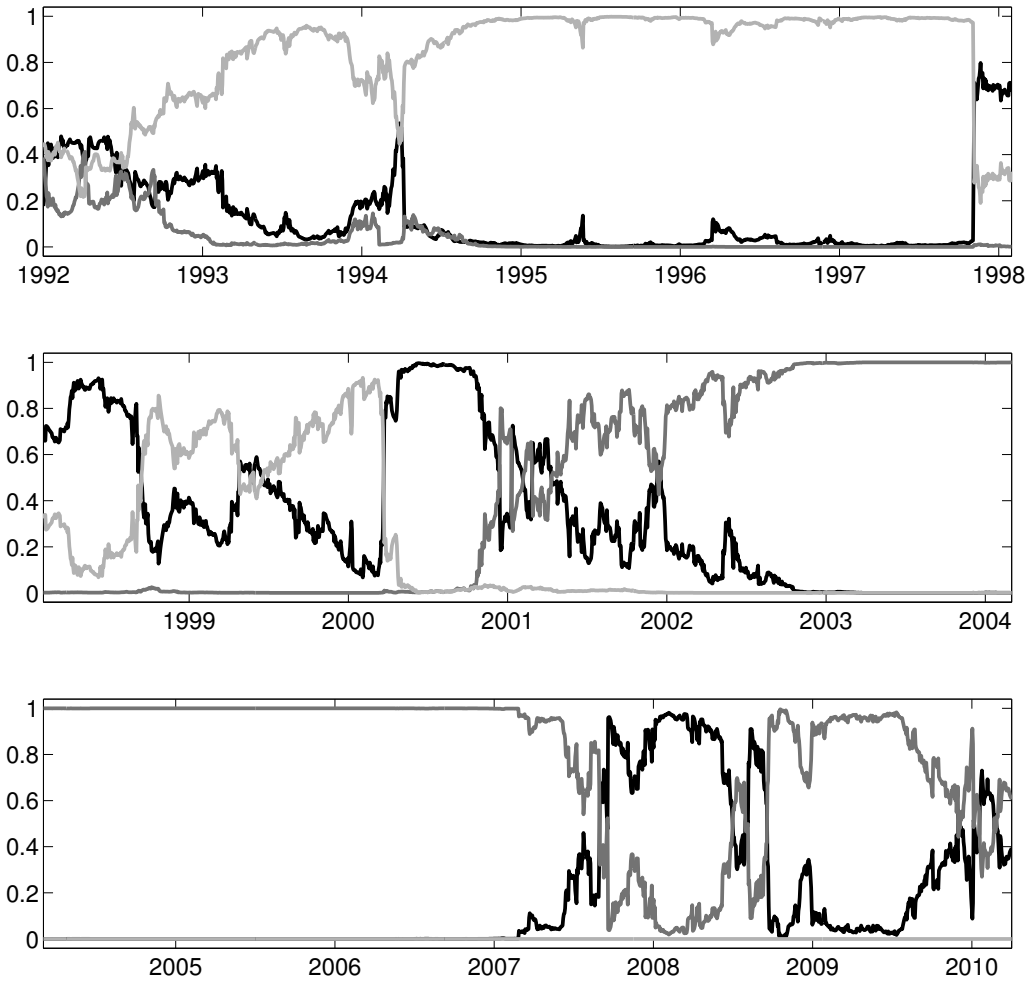

Figure 6: Bayesian model averaging weights, updated each trading day: sum of weights for daily models (black), high frequency models (dark grey), and options models (light grey).

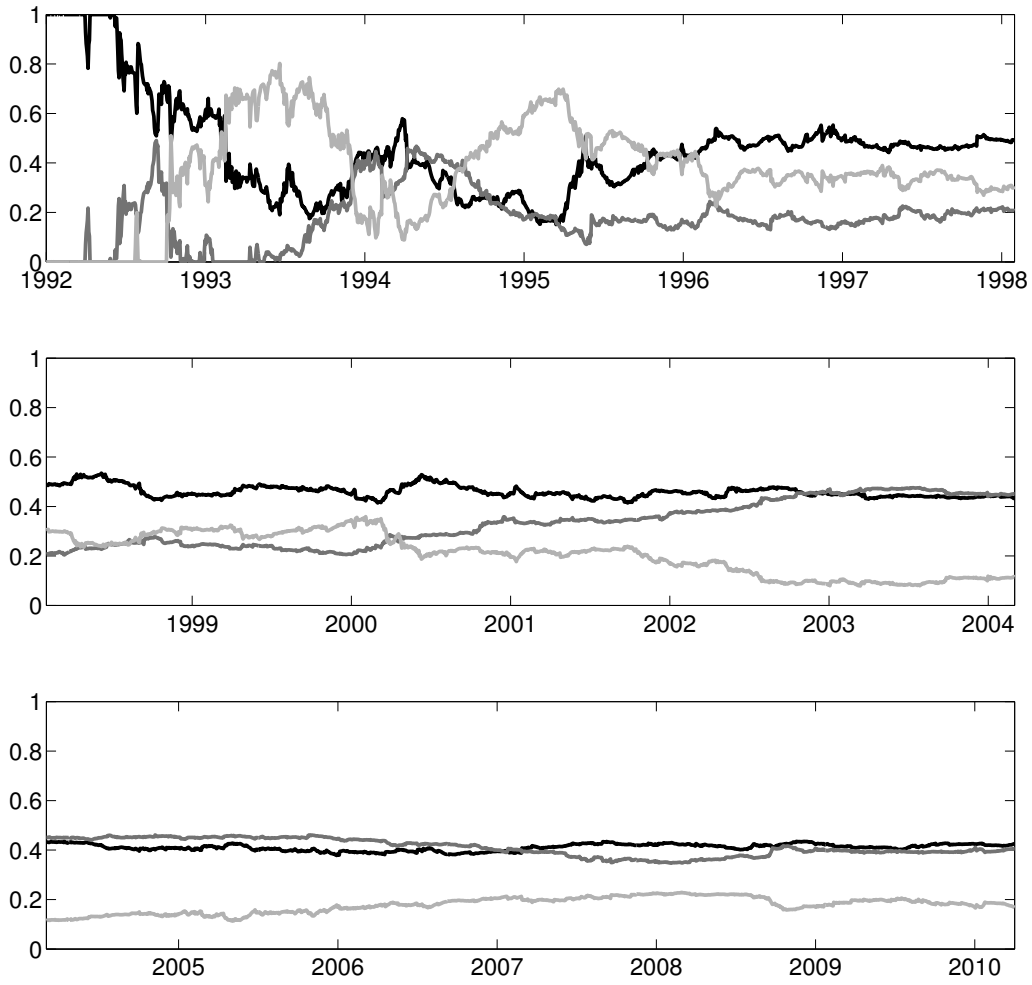

Figure 7: Optimal prediction pool weights, updated each trading day: sum of weights for daily models (black), high frequency models (dark grey), and options models (light grey).

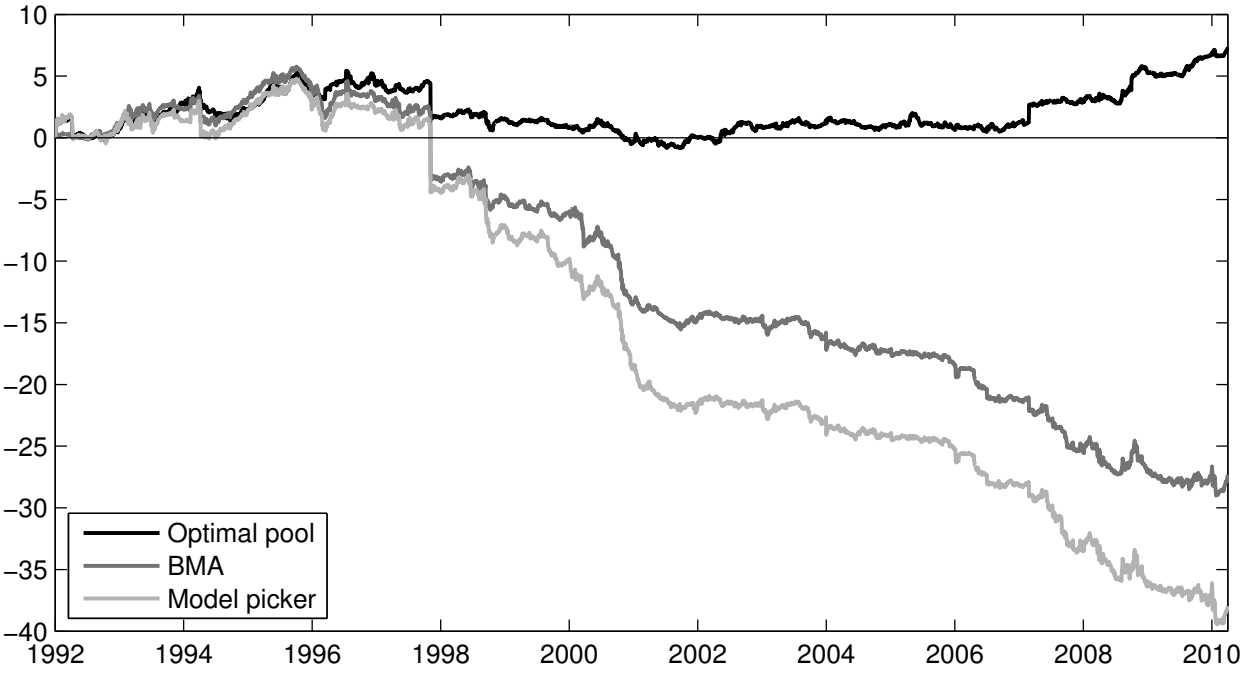

Figure 8: Log scores, differences relative to equally-weighted pool.

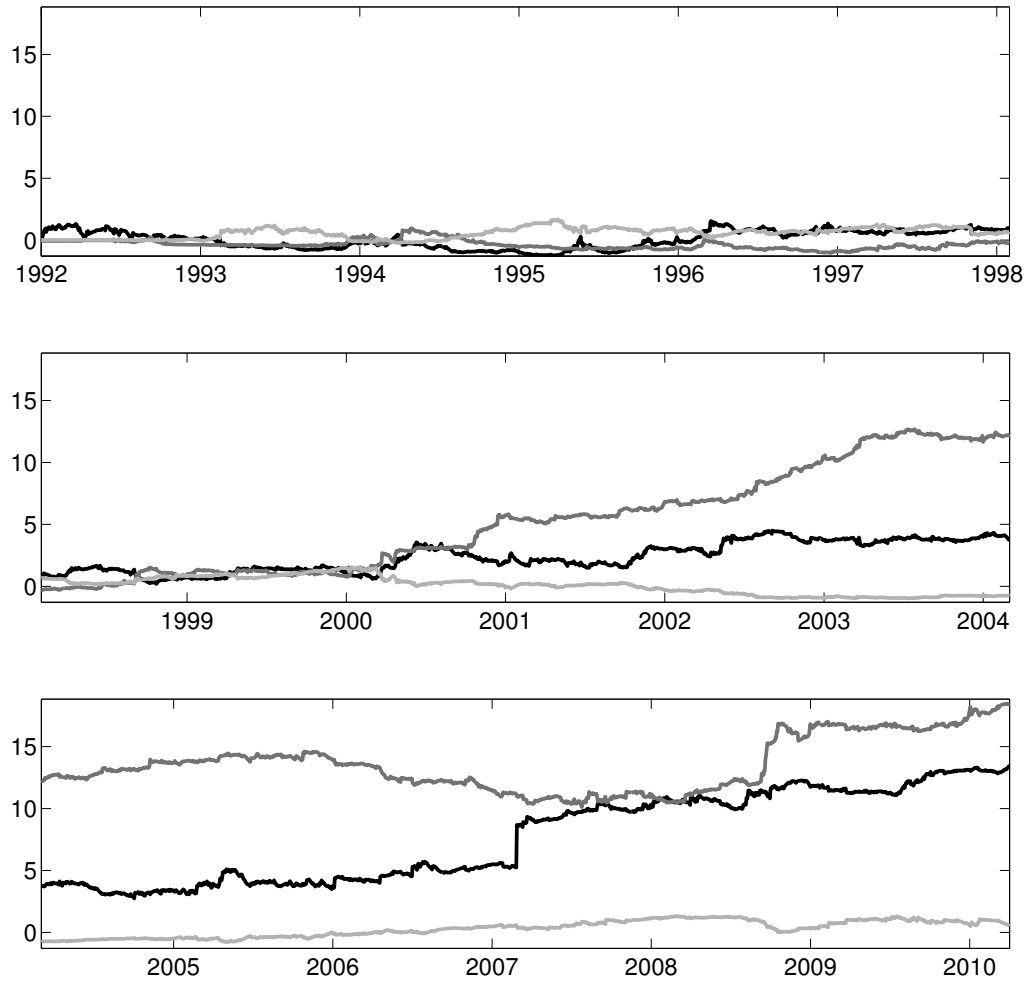

Figure 9: Values of the group of daily models (black), high frequency models (dark grey), and options models (light grey).